Phase Identification with Incomplete Data

Minghao Xu, Ran Li, and Furong Li, *Senior Member, IEEE*

***Abstract*— Phase identification is a process to determine which of the three phases a particular house is connected to. The state-of–the-art identification methods usually exploit smart metering data. However, the data sets are not always available and the major challenge is hence to identify phases with incomplete data set. This paper proposes a novel spectral and saliency analysis (SSA) identification method to overcome this hurdle. Spectral analysis is firstly performed to extract the high-frequency features from the incomplete data. Saliency analysis is then adopted to extract salient features from the variations of high-frequency loads in the time domain. Correlation analysis between customer features and the phase features is used to determine customers' phase connectivity. The method is executed in an iterative manner until all the consumers are identified or no salient features can be found. It is validated against real data from over 6000 smart meters in Ireland and achieves an accuracy of over 93% with only 10% smart meter penetration ratio in a 100 household network.**

***Index Terms*—Phase identification, spectral analysis, LV distribution network, smart metering data, incomplete data set.**

I. Introduction

Traditional research and design on low voltage (LV) distribution level rarely take the phase connectivity of individual consumption into consideration [1]. This leads to an urgent problem that the existing networks are poorly 3-phase balanced [2]. Such unbalanced loads will lead to extra power loss and reduced lifespans of assets. Recently in the UK, the 3-phase imbalance issue in the existing LV networks have been further aggravated due to the wide deployment of Low Carbon Technologies (LCTs) at household level. In order to accommodate the fast growing LCTs meanwhile considering the phase balance of LV networks, a vital problem that the Distribution Network Operators (DNOs) are facing is to identify which phase a particular house is connected to.

Traditionally, the DNOs would send electricians to check the phase connectivity manually in the field which is inherently inefficient. Installing advanced signal injecting and receiving equipment on both ends of the networks [3, 4] is another option for the DNOs. These devices are accurate and fast but are at the cost of increased capital and maintenance fees. The introduction of other high-precision metering devices [5-9] provides another opportunity to identify phases in an indirect way. The inevitable high cost of these devices become the main obstacle for them to be widely deployed.

Recently, with the roll-out of smart meters, data-driven methods based on the analysis of smarting data have been developed. By the data type they require, these methods could be categorized into two sets:

1) Voltage data [7, 8, 10, 11]: measuring shape similarities between household voltage and phase voltage measured at substation through correlation analysis, regression or clustering techniques. It assumes that consumers share similar voltage patterns within the same phase;

2) Load data [12-15]: based on the law of conservation of energy, finding the optimal combination of households to provide similar aggregation load as the phase load.

However, the limitation of the first category of methods is that voltage data are not commonly provided by most smart meters [16]. In the second category, the methods are designed for handling data with small degrees of loss or error, requiring that the distribution networks to be analyzed should have 100% or nearly 100% penetration ratios of smart meters. Whereas smart meters are not widely deployed in most places. In the UK consumers could even opt not to install smart meters [17]. Therefore, there is a critical need to develop phase identification methods with incomplete consumer data i.e., to perform phase identification with only a proportion of consumers having smart meters in the network.

In this paper, a novel approach based on spectral and saliency analysis (SSA) of consumers' load has been developed. SSA aims to extract customer's load features from both time and frequency domains. Hence it could effectively identify phase connection from limited data compared with traditional methods which directly operate on the raw data. Firstly, spectral analysis using Fourier Transform on both consumer's load data and phase load data is performed to filter out the low-frequency components. Then the variations of each consumer's remaining high-frequency load are extracted as their features. Following that, the saliency of these features are assessed to form the salient feature vectors for consumers. Lastly, the salient feature vectors of each consumer are correlated with the corresponding high-frequency variations on each of the three phases. Given the load variations in salient feature vector are significant, the corresponding phase variations should present similar variation pattern and the phase can therefore be identified. To the best of the users' knowledge, this is the first time phase identification has been achieved under incomplete load data condition. It is validated using real smart metering data form Smart Meter Trial in

M. Xu, R. Li and F. Li are with the Department of Electronic and Electrical Engineering, University of Bath, Bath BA2 7AY, UK (e-mail: mx266@bath.ac.uk, rl272@bath.ac.uk, f.li@bath.ac.uk)

Ireland [18]. This paper has evaluated the accuracies of proposed phase identification method under different data conditions. It was carried out by gradually changing the smart meter penetration ratio and time length of the available data.

The reminder of this paper is organized as follows: Section II shows how the problem is mathematically formulated and Section III presents the proposed method. Section IV validates the proposed approach and compares it with other technique. Section V draws the conclusions.

II. Problem Formulation

The mathematical model for the problem is developed as follow.

Suppose there are n consumers in this network and for each consumer, m measurements of load data are taken by smart meters over the time. Since the network is three-phase, the set of indices of phases, J, consists only three elements. (1), (2) and (3) represent sets of the indices for phases, consumers, and measurements respectively.

$$J = \{1, 2, 3\} \tag{1}$$

$$C = \{1, 2, \dots, n\} \tag{2}$$

$$M = \{1, 2, \dots, m\} \tag{3}$$

Let h_{ki} represents the measured load at time k for consumer i. p_{kp} denotes the substation's load for phase p at time k. Consumer load matrix H and phase load matrix L are expressed in (4) (5) respectively.

$$H = \begin{bmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{bmatrix} \quad \forall n \in C \quad \forall m \in M \tag{4}$$

$$L = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ \vdots & \vdots & \vdots \\ p_{k1} & p_{k2} & p_{k3} \\ \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & p_{m3} \end{bmatrix} \quad \forall k \in M \tag{5}$$

Let x_{ij} be the phase indicator of consumer i to phase j. 1 means true and 0 indicates false. The connectivity matrix would then be as follows.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & x_{i1} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix} \tag{6}$$

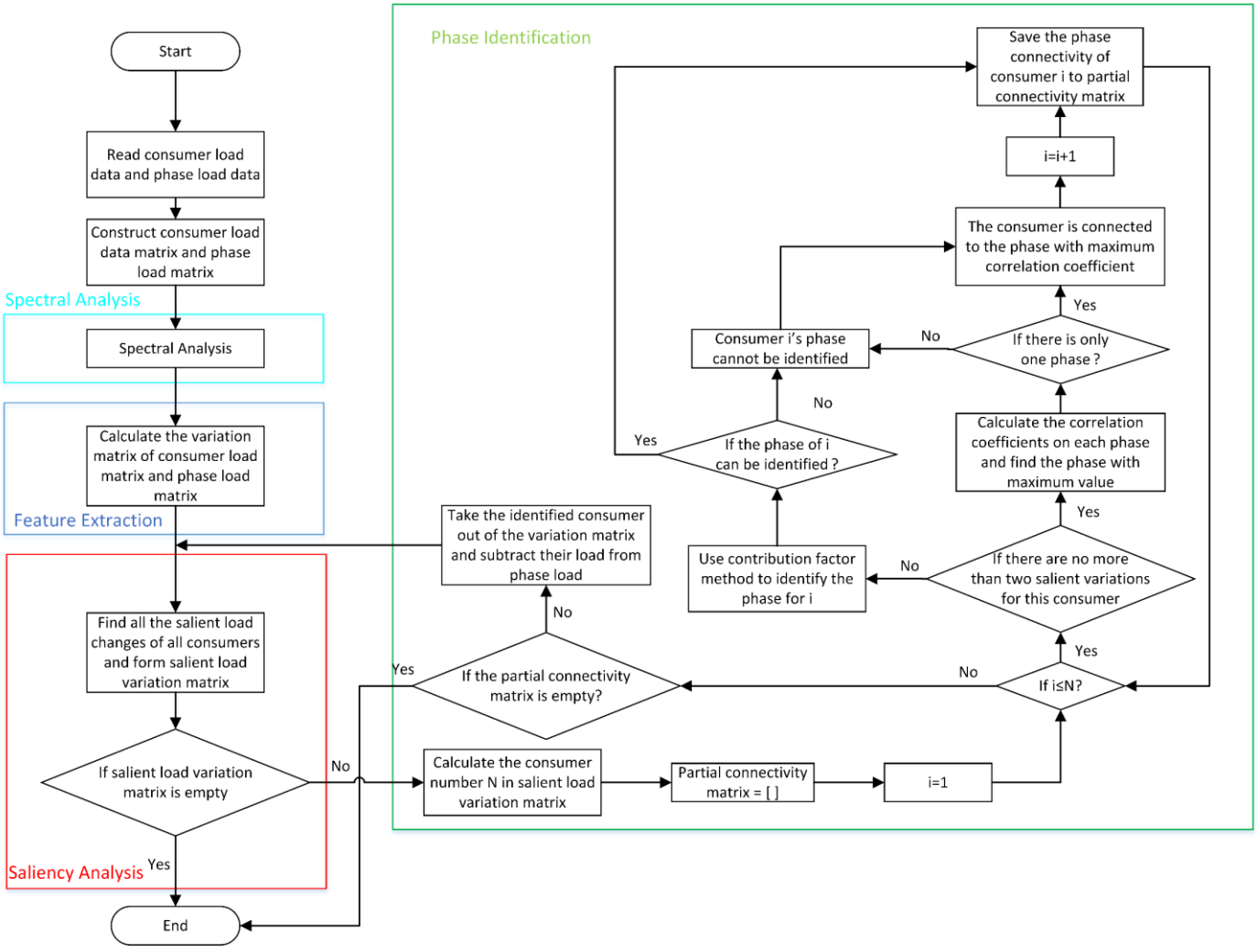

Fig. 1. Flowchart for phase identification using saliency analysis algorithm

Traditionally, the problem is formulated as (7) and various optimization techniques could be applied to get the optimal mathematical solution of X.

$$HX = L \tag{7}$$

However, due to the above formulation, all of the existing methods are fundamentally derived from the law of conservation of energy, i.e., the sum of the individual's loads within the same phase is equal to the corresponding phase load monitored at the substation. The inevitable problem caused by this is that under incomplete data condition, the accuracy of X would be fairly poor.

III. Spectral and Saliency Analysis Algorithm

To tackle the limitation under incomplete data condition, this paper proposes to extract distinct features from individual load profiles and correlate it with phase load to estimate the connectivity. Fig 1 is the overall flowchart for this algorithm. It is achieved by executing the following several steps.

Firstly, it filters out the low-frequency load from each household's and phase load by performing Fourier Transform (FT) and Inverse Fourier Transform (IFT). Then it takes the variations of the remaining load between two arbitrary time intervals as the features. After that, the algorithm analyses the saliency of all the features and extracts the salient variations of each consumer. Lastly, depending on the number of salient variations of the consumer, the method identifies the phase by either correlation analysis or the contribution factor analysis proposed in this paper. The algorithm then removes the identified consumers from the overall data set and the above steps will be repeated until: 1) there are no salient variations can be found; 2) no consumer's phase can be identified by analyzing the selected salient variations. Detailed explanations are presented in the following subsections.

A. Spectral Analysis

The first step is to perform spectral analysis on both the consumers' load data and the phase load data. Discrete Fourier Transform (DFT) is applied in the paper to get the spectrum of the data. Suppose b is a time series load profile, which in this paper represents an arbitrary column in the consumer load matrix H or an arbitrary column in the phase load matrix L. The DFT of b is formulated below.

$$B_k = \sum_{n=0}^{m-1} b_n e^{-j\frac{2\pi kn}{m}}, k = 0, \dots, m-1 \tag{8}$$

where m is the number of measurement and $B_k = \beta_k e^{j\theta_k}$ is the frequency spectrum with magnitude β_k and phase angle θ_k.

The low-frequency components of load data mostly follow a regular pattern while high-frequency components are different from house to house, representing the unique energy usage habit of the customer. After setting a cut-off frequency, f_c, the magnitudes of the low-frequency harmonics, whose frequencies are below f_c, are all set to zeros. Due to the symmetry property of DFT, the harmonics that are symmetrical with the low-frequency harmonics about the Nyquist frequency should be set to zeros as well to completely filter out the low-frequency harmonics. Then the high-frequency load profile in time domain can be obtained by applying Inverse Discrete Fourier Transform (IDFT) with the remaining frequency spectrum.

$$b_n^r = \frac{1}{m}\sum_{k=0}^{m-1} B_k e^{j\frac{2\pi kn}{m}}, n = 0, \dots, m-1 \tag{9}$$

where b^r is the reconstructed time series load profile.

After performing DFT and IDFT on the consumer load matrix H and phase load matrix L, the high-frequency parts of the consumer load H_{high} and phase load L_{high} are obtained.

$$H_{high} = \begin{bmatrix} h_{high11} & \cdots & h_{high1n} \\ \vdots & \ddots & \vdots \\ h_{highm1} & \cdots & h_{highmn} \end{bmatrix} \quad \forall n \in C \quad \forall m \in M \tag{10}$$

$$L_{high} = \begin{bmatrix} p_{high11} & p_{high12} & p_{high13} \\ \vdots & \vdots & \vdots \\ p_{highk1} & p_{highk2} & p_{highk3} \\ \vdots & \vdots & \vdots \\ p_{highm1} & p_{highm2} & p_{highm3} \end{bmatrix} \quad \forall k \in M \tag{11}$$

B. Feature Extraction

The second step is to extract the features from the remaining high-frequency load profiles. As mentioned in the introduction section, there exist identification methods using signal injecting equipment. The injector poses a unique electric signal from the demand side. At the substation, there are three receivers waiting to detect the signal. The phase at which the receiver captures the injected signal is the phase which the household is connected to. This is in nature to detect the external turbulence in the network. Similarly, the saliency analysis method is proposed in this paper. Instead of injecting external signals into the network, the proposed method seeks salient high-frequency load variations of consumers. For example, during a period, all the households are consuming electricity at a steady level, except for one household. The residents living in this house may return home late and turn on the light, kettle, etc. As a result, the demand or consumed energy within this period for this particular household would increase significantly. On the substation side, the corresponding phase load would increase accordingly. Since other households in the network are consuming energy almost constantly during this period, the load increase at the corresponding phase should be quite noticeable. The phase of the household could then be identified. In this paper, the high-frequency loads are obtained to reveal the unique energy usage habit of the households, and the variations of the high-frequency loads are extracted as features.

The variation of consumer i between two adjacent time intervals k and $(k+1)$ are calculated by (12). They can reflect the change of consumer's energy behavior.

$$Vh_{hki} = h_{high(k+1)i} - h_{highki} \quad \forall k \in M \quad k \neq 1 \quad \forall i \in C \tag{12}$$

The variation of phase j between periods k and $(k+1)$ is expressed as (13).

$$Vp_{kj} = p_{high(k+1)j} - p_{highkj} \quad \forall k \in M \quad k \neq 1 \quad \forall j \in J \tag{13}$$

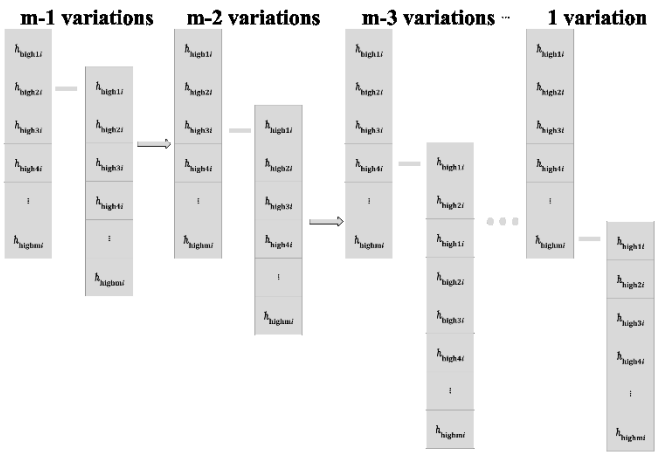

Fig. 2. Load variations between two arbitrary time intervals

Hence, the variation matrices of consumer load and phase load are shown in (14) and (15) respectively.

$$VH_1 = \begin{bmatrix} Vh_{11} & \cdots & Vh_{1n} \\ \vdots & \ddots & \vdots \\ Vh_{(m-1)1} & \cdots & Vh_{(m-1)n} \end{bmatrix} \tag{14}$$

$$VL_1 = \begin{bmatrix} Vp_{11} & Vp_{12} & Vp_{13} \\ \vdots & \vdots & \vdots \\ Vp_{k1} & Vp_{k2} & Vp_{k3} \\ \vdots & \vdots & \vdots \\ Vp_{(m-1)1} & Vp_{(m-1)2} & Vp_{(m-1)3} \end{bmatrix} \tag{15}$$

The above variation matrices only represent the variations between two adjacent time intervals. The variation matrices could be further expanded by introducing variations between any two time intervals. Fig 2 demonstrates all the possible load variations of consumer i. There are $m-1$ pairs of columns and each column represents the consumer's load data with m measurements. The right-hand-side column in each pair slides downwards gradually. By subtracting the right-hand-side data from the corresponding left-hand-side data, all the variations can be calculated. For consumer with m measured loads, the total number of variations, N_{va}, is given by (16).

$$\begin{aligned} N_{va} &= (m-1) + (m-2) + (m-3) + \cdots + 1 \\ &= (m-1)m/2 \qquad\qquad m > 1 \end{aligned} \tag{16}$$

The variation of consumer i's load between time interval k and $(k+\mathrm{t})$ can then be expressed as (17).

$$\begin{aligned} Vh_{kit} &= h_{high(k+t)i} - h_{high(ki)} \\ \forall t \in M \quad t \neq m \quad \forall k \in M \quad k \neq 1 \quad \forall i \in C \end{aligned} \tag{17}$$

The load variation of phase j between periods k and $(k+$

$$\begin{aligned} Vp_{kjt} &= p_{high(k+t)j} - p_{high(kj)} \\ \forall t \in M \quad t \neq m \quad \forall k \in M \quad k \neq 1 \quad \forall j \in J \end{aligned} \tag{18}$$

The variation matrices of consumer load then become (19).

$$VH = \begin{bmatrix} VH_1 \\ \vdots \\ VH_t \\ \vdots \\ VH_{m-1} \end{bmatrix} \quad \forall t \in M \quad t \neq m \tag{19}$$

where VH_t represents

$$VH_t = \begin{bmatrix} Vh_{11t} & \cdots & Vh_{1nt} \\ \vdots & \ddots & \vdots \\ Vh_{(m-t)1t} & \cdots & Vh_{(m-t)nt} \end{bmatrix} \tag{20}$$

The variation matrix of phase load is given in (21).

$$VL = \begin{bmatrix} VL_1 \\ \vdots \\ VL_t \\ \vdots \\ VL_{m-1} \end{bmatrix} \quad \forall t \in M \quad t \neq m \tag{21}$$

where VL_t represents

$$VL_t = \begin{bmatrix} Vp_{11t} & Vp_{12t} & Vp_{13t} \\ \vdots & \vdots & \vdots \\ Vp_{k1t} & Vp_{k2t} & Vp_{k3t} \\ \vdots & \vdots & \vdots \\ Vp_{(m-t)1t} & Vp_{(m-t)2t} & Vp_{(m-t)3t} \end{bmatrix} \tag{22}$$
$$\forall t \in M \quad t \neq m \quad k \in M \quad k \leq m-t$$

C. Saliency Analysis

After extracting the variations as the feature, the next step is to analyze the features, i.e., to identify the salient variations. Within the same time step, if a consumer's load variation is significantly higher than the sum of other loads' variation, this load variation is defined as a salient variation. Mathematically, the salient changes are defined as the changes which satisfies the following condition.

$$|Vh_{kit}| \geq TH \times \sum_{c=1, c\neq k}^{n} Vh_{kct} \tag{23}$$

where:

Vh_{kit}	The load change of consumer i during time interval $k+t$ and k;
n	The consumer number in the network;
TH	The threshold value to adjust how salient the changes are.

The reason for selecting the salient variations is that they are more likely to be observed from phase load variation. In other words, since each consumer is connected to only one phase, the salient variation of one consumer is more likely to cause the load variation of the corresponding phase. TH can be changed so that the saliency is adjustable according to various data conditions.

D. Phase Identification

For each consumer with m measurements, there are in total $(m-1)m/2$ variations could be calculated. Suppose there are g salient variations for consumer i. There salient variations form a row vector SVh_i.

$$SVh_i = [SVh_{i1} \quad SVh_{i2} \quad \cdots \quad SVh_{ig}] \tag{24}$$

Accordingly, l phase load variations can be found for each phase, SVl_{i1}, SVl_{i2}, SVl_{i3}.

$$SVl_{i1} = [SVl_{i11} \quad SVl_{i12} \quad \cdots \quad SVl_{i1g}] \tag{25}$$

$$SVl_{i2} = [SVl_{i21} \quad SVl_{i12} \quad \cdots \quad SVl_{i1g}] \tag{26}$$

$$SVl_{i3} = [SVl_{i31} \quad SVl_{i32} \quad \cdots \quad SVl_{i3g}] \tag{27}$$

The consumer's salient variations are correlated with the corresponding phase variations on each of the three phases. Three correlation coefficients are then obtained for each consumer. By selecting the phase which is tightly coupled with the consumer's load variations, the phase connectivity of the consumer can then be identified. Pearson's correlation coefficient is adopted to indicate how strong the consumer's load variation and the phase load variation are correlated with each other. It is formulated below.

$$\rho\left(SVh_i, SVl_{ij}\right) = \frac{cov\left(SVh_i, SVl_{ij}\right)}{\sigma_{SVh_i}\,\sigma_{SVl_{ij}}} = \frac{1}{g-1}\sum_{ind=1}^{g}\frac{(SVh_{i\,ind} - \mu_{SVh_i})(SVl_{ij\,ind} - \mu_{SVl_{ij}})}{\sigma_{SVh_i}\sigma_{SVl_{ij}}} \quad (28)$$

where:

SVh_i	Salient variations of consumer i;
SVl_{ij}	Corresponding salient variations of consumer i on phase j;
$cov(SVh_i, SVl_{ij})$	Covariance of SVh_i and SVl_{ij};
σ_{SVh_i}	Standard deviation of SVh_i;
$\sigma_{SVl_{ij}}$	Standard deviation of SVl_{ij}.
g	Number of salient variations for consumer i, $g > 1$;
ind	Index of salient variation in the row vector;
μ_{SVh_i}	The mean value of SVh_i;
$\mu_{SVl_{ij}}$	The mean value of SVl_{ij}.

ρ is a quantitative measure of the correlation between two series. After computing the three coefficients for each consumer, phase of the consumer is identified by selecting the phase with maximum correlation coefficient.

However, when the load data are not measured for a long period (m is small), the number of salient variations (g) of the consumer is likely to be small. When there is only one salient variation, $(g-1)$ and the two standard deviations equal to zero. As a result, (28) cannot be used. As for consumers with two salient variation, the correlation coefficients could only reflect the changing trend of the two series. In other words, whatever the two variations are in the two series, ρ will be either 1 or -1. This is proved as follows.

Suppose the two salient variations for consumer i are as (29).

$$SVh_i = [SVh_{i1} \quad SVh_{i2}] \quad (29)$$

The corresponding load variation on phase j is as follows.

$$SVl_{ij} = [SVl_{ij1} \quad SVl_{ij2}] \quad (30)$$

Substitute the variables in (29), (31) is obtained.

$$\rho\left(SVh_i, SVl_{ij}\right) = \frac{(SVh_{i1} - SVh_{i2})(SVl_{ij1} - SVl_{ij2})}{|SVh_{i1} - SVh_{i2}||SVl_{ij1} - SVl_{ij2}|} \quad (31)$$

For each consumer with only two salient variations, three ρ values can be obtained. While as can be seen from the above equation, the ρ value will be only 1 or -1. This does not provide enough information for phase identification. To tackle the problem caused by consumers with no more than two salient variations, the contribution factor is defined as follows.

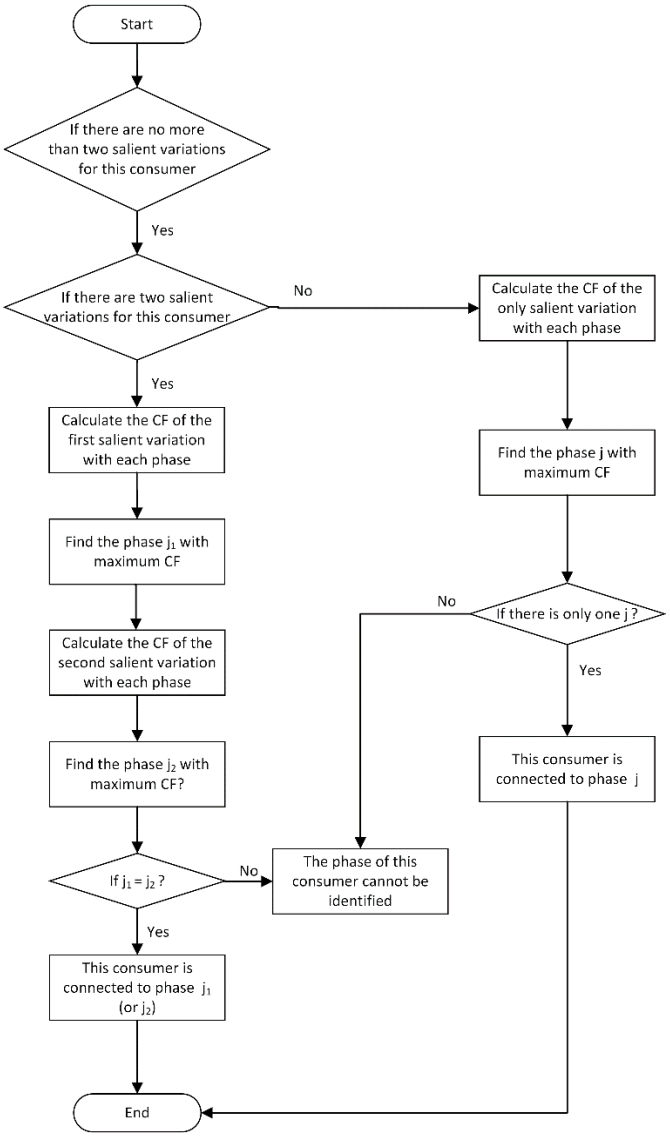

Fig. 3. Phase identification flow chart for consumers with no more two salient variations

$$CF_{i\,j\,ind} = \frac{SVh_{i\,ind}}{SVl_{ij\,ind}} \quad (32)$$

where:

ind	Index of salient variations of consumer i in the row vector;
j	Phase index, $\forall j \in J$
$CF_{i\,j\,ind}$	The contribution factor of the ind's salient variation of consumer i to the corresponding load variation on phase j;

The ratio, $CF_{i\,j\,ind}$, is a measure of how much the phase

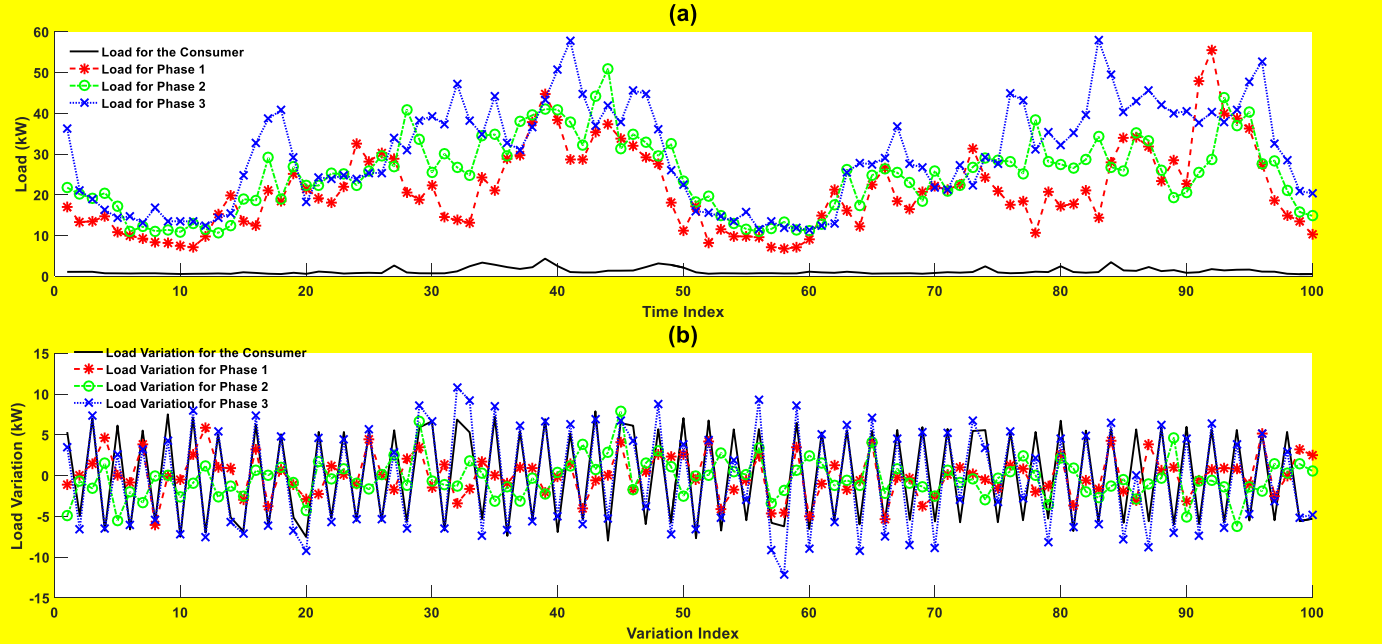

Fig. 4. Correlation improvement: (a) Original load for the consumer and the three phases; (b) Salient variations of the consumer and the three phase variations during the corresponding intervals.

variation is contributed by the consumer's load variation. With larger value of $CF_{i\ j\ ind}$, there is more confidence in estimating that the consumer is connected to this phase. The contribution factor could be used to help identify phase connectivity for consumer with no than two salient variations. Fig 3 shows how phase is identified for a consumer with no more than two salient variations. This is part of the overall algorithm shown in Fig 2.

IV. Validation and Results

The proposed approach is validated using real smart metering data from the Smart Metering Electricity Customer Behaviour Trials (CBTs) initiated by Commission for Energy Regulation (CER) in Ireland [18]. The data were taken on a half-hourly basis from 1st July 2009 to 31st December 2010.

A. Improvement of Correlation

Unlike voltage data, load data are highly dependent on consumers' behaviours and they do not share any similar patterns within the same phase. In this part, the test shows how the individual consumption correlates with its phase load and how the proposed SSA method improves the degree of correlation.

In this preliminary test, 90 consumers out of 100 in the network have installed smart meters. A total length of two-month data were used to perform phase identification. The black line in Fig 4 (a) shows the load variations over 100 consecutive intervals of a randomly selected consumer connected to phase 3 in the network. As for the red, green, and blue curves, they represent the load on phase 1, phase 2 and phase 3 respectively. As can be observed, both the amplitude and the shape of the curves vary a lot from each other. No noticeable correlated relationship can be observed.

Fig 4 (b) takes the same consumer as in Fig 4 (a) but extracts the consumer's salient variations after spectral analysis as the salient feature vector, represented by the black curve. The red, green, and blue lines demonstrate the load variations during the same periods as the salient load variations on phase 1, phase 2 and phase 3 respectively. Phase 3, which is in blue, is obviously correlated with the consumer's salient load variations. This intuitively indicates that the consumer is connected to phase 3 which matches the real case.

TABLE I
CORRELATION COEFFICIENTS BETWEEN CONSUMER AND PHASES

Type	Phase Number	Correlation Coefficients
Correlation analysis of original data	Phase 1	0. 1864
	Phase 2	0. 1662
	Phase 3	0. 1773
Correlation analysis after SSA	Phase 1	-0.1020
	Phase 2	-0.4662
	Phase 3	0.9301

Table I gives the Pearson correlation coefficients between the consumer and the three phases. The original load of selected consumer and phase 3 are only correlated by a coefficient of 0.1773, which is even lower than the coefficient with phase 1. However, after SSA, the correlation coefficient is significantly improved to 0.9301, which is much higher than the other two phases.

B. Performance Evaluation of the Proposed Method

This section takes the same distribution network as in Section A which consists of 100 domestic consumers. A comprehensive evaluation is performed by gradually adjusting the smart meter penetration ratio and length of data to be used.

The results of the evaluation are presented in Table II. As shown in the table, the penetration ratio of smart meters in the network increases from 10% to 100% along the horizontal direction and the data length increase from 1 month to 12 months along the vertical direction. Under each data condition, phase identification is performed for multiple times to get an average accuracy. It can be observed in the table that all the

TABLE II
OVERALL IDENTIFICATION ACCURACY

Time length (month)	Smart meter penetration ratios in the network									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	93.29%	96.33%	96.00%	96.84%	98.16%	99.04%	99.96%	99.63%	100.00%	100.00%
2	95.32%	97.35%	97.35%	99.88%	100.00%	99.71%	98.80%	100.00%	100.00%	100.00%
3	99.38%	97.35%	99.38%	98.87%	99.78%	98.70%	100.00%	100.00%	100.00%	100.00%
4	97.35%	96.33%	97.35%	98.36%	98.57%	98.02%	100.00%	100.00%	100.00%	100.00%
5	100.00%	100.00%	96.00%	97.86%	100.00%	99.04%	99.38%	100.00%	100.00%	100.00%
6	97.35%	96.33%	96.67%	99.38%	98.16%	98.70%	100.00%	100.00%	99.60%	100.00%
7	99.38%	100.00%	96.67%	99.38%	98.97%	98.70%	100.00%	100.00%	100.00%	100.00%
8	100.00%	95.32%	98.70%	98.87%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
9	97.35%	96.33%	96.00%	99.38%	96.94%	98.70%	100.00%	99.88%	100.00%	100.00%
10	100.00%	98.36%	100.00%	99.38%	100.00%	100.00%	99.96%	100.00%	100.00%	100.00%
11	95.32%	97.35%	97.35%	100.00%	98.16%	99.71%	99.38%	100.00%	100.00%	100.00%
12	100.00%	99.38%	98.02%	99.88%	99.38%	99.04%	99.67%	100.00%	100.00%	100.00%

identification accuracies are satisfactory and are all above 95% except the most upper-left condition which is identification with one month data and with 10% smart meter penetration ratio. With fixed data length, the general trend is that the identification accuracy would increase gradually with the increase in smart meter penetration ratio. The reason is that during saliency analysis, each household's high-frequency loads has to be compared with all others' high-frequency loads. With higher smart meter penetration ratio, the saliency analysis becomes more thorough and will return results with more precision and confidence.

C. Comparison with Other Published Method

To our best knowledge, the most comparable method to the proposed algorithm is [12], where the identification process is formulated as a Mixed Integer Quadratic Programming (MIQP) problem. However, this method is designed for handling data with small degrees of loss or error. Additionally, the MIQP method was validated using artificially generated load data and has never been tested under real incomplete data condition. In the paper, the authors replicated the work with small modifications by considering the missing data as noise. Due to license issue, the optimization solver used in this paper is Gurobi instead of CPLEX which was used in the original paper.

To demonstrate the significance of the proposed method over the MIQP method in larger networks, the comparison are performed in 200, 400, 600, 800, and 1000 households networks under various data condition. The results are in Fig. 5. In the figure, there are in total 10 rectangles in two rows

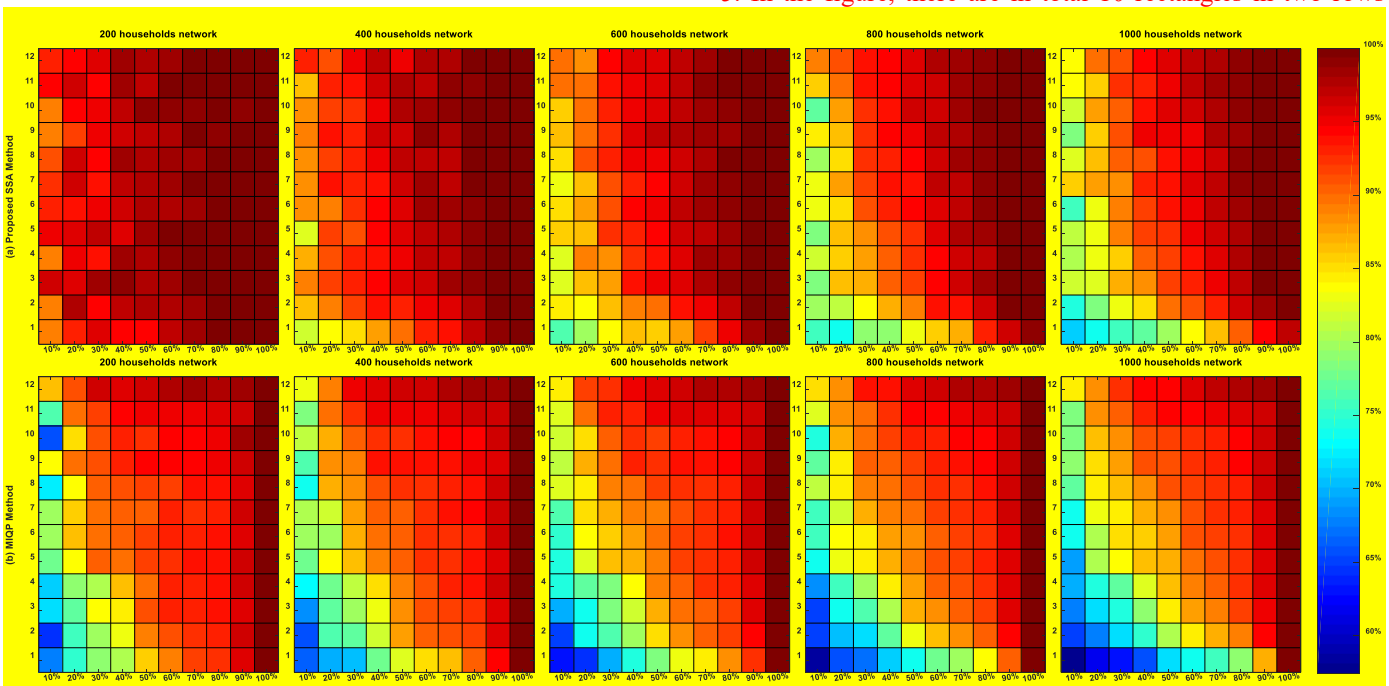

Fig. 5. Comparison of the Performance of the proposed SSA method and MIQP Method: (a) Overall identified accuracies vary with used time and penetration ratios of smart meters using the proposed SSA method in 200, 400, 600, 800, and 1000 households networks; (b) Overall identified accuracies vary with used time and penetration ratios of smart meters using MIQP method in 200, 400, 600, 800, and 1000 households networks.

wherein the upper row represents the results using the proposed SSA method and the lower row represents the results using MIQP method. Each rectangle consists of 120 (12 × 10) colour coded sub-rectangles and is a graphical representation of the identification accuracy. Within each rectangle, the vertical direction indicates the data length used (1-12 months) and the horizontal direction shows the penetration ratio of smart meters ranging (from 10% to 100%). All of the ten rectangles share a unified colour bar as shown on the right-hand side in Fig. 5. The colour varies from deep blue to deep red with deep red representing 100% identification accuracy.

As illustrated in the figure, several findings are:

1) Both of the methods provides 100% accuracies under complete data condition, i.e., the smart meter penetration ratio is 100%;
2) The performance of both of the methods decrease as the smart meter penetration ratio decreases or the data length decreases;
3) With the increase in network size, both of the methods tend to become less accurate slightly. For SSA, the reason is that, with larger network size, the salient features of individual households are more likely to cancel out with each other through aggregation on the phase load and hence it becomes more difficult to find these salient features. For MIQP, with increased network size, the number of possible combinations of households to provide similar aggregation load would increase significantly and hence the accuracy would drop accordingly;
4) The proposed SSA method outperforms the MIQP method under almost every scenario, especially when the data is incomplete (low smart meter penetration and short data length). This can be observed in the top left of every rectangle in Fig. 5 as the upper row are much 'redder' than the ones in the lower row.

To give a clear comparison, the identification accuracy with 10% smart meter penetration ratios in each rectangle are calculated are given in Table III. Compared with the MIQP method, the proposed method lifts the accuracy for networks with 200, 400, 600, 800, and 1000 households by 21.08%, 15.97%, 13.69%, 10.78%, and 9.96% respectively.

TABLE III
COMPARISON OF IDENTIFICATION ACCURACIES WITH 10% SMART METER PENETRATION RATIO

Households Number	Average Accuracy		Accuracy Improvement Percentage
	MIQP Method	Proposed Method	
200	75.87%	91.86%	21.08%
400	75.27%	87.29%	15.97%
600	74.74%	84.97%	13.69%
800	73.46%	81.38%	10.78%
1000	72.79%	80.04%	9.96%

V. CONCLUSION

This paper proposes a novel phase identification method based on spectral and saliency analysis which can be used under incomplete data conditions. It essentially changes the rule on which most of the current identification methods are based. Also, it successfully develops a contribution factor and introduces correlation coefficient to help identify phases after SSA.

To demonstrate the significance of the method, an LV distribution network of 100 domestic consumers has been constructed using real smart metering data from Ireland to perform a preliminary test. Subsequent to the test, a comprehensive and thorough evaluation of the proposed method has been undertaken. Additionally, the performance of the proposed method has been compared with the available optimization method. Results have shown that:

- The SSA could help reveal the correlation between the consumer and the corresponding phase;
- The identification performance of the proposed method would grow with the increase in smart meter penetration ratio and load data length;
- The identification performance of the proposed method would decrease slightly with the increase in network size, i.e., the increase in the total number of household in the network;
- The proposed method outperforms the available method in almost every aspect. Under the extreme data condition where the smart meter penetration ratio is 10% in a 200-household network, the proposed method achieved an average identification accuracy of 91.86 % which is 21.08% higher than the available MIQP method.

The proposed method cannot only reduce uncertainties during the development of smart grid, but also provide necessary tool for DNOs to balance current networks.

Future work will focus on dynamic salience analysis, i.e., saliency criteria will adjust according to different data conditions. Additionally, the proposed method will be validated using networks with multiple consumer types.

REFERENCES

[1] H. L. Willis, "Characteristics of Distribution Loads," in *Electrical Transmission & Distribution Reference Book*, ed: ABB Power T&D Company Inc., 1997, pp. 784-808.
[2] K. Ma, R. Li, and F. Li, "Quantification of Additional Asset Reinforcement Cost From 3-Phase Imbalance," *IEEE Transactions on Power Systems,* vol. PP, pp. 1-7, 2015.
[3] K. J. Caird, "Meter phase identification," ed: US Patents, 2010.
[4] S. Zhiyu, M. Jaksic, P. Mattavelli, D. Boroyevich, J. Verhulst, and M. Belkhayat, "Three-phase AC system impedance measurement unit (IMU) using chirp signal injection," in *Applied Power Electronics Conference and Exposition (APEC), 2013 Twenty-Eighth Annual IEEE*, 2013, pp. 2666-2673.
[5] C. Chao-Shun, K. Te-Tien, and L. Chia-Hung, "Design of Phase Identification System to Support Three-Phase Loading Balance of Distribution Feeders," *Industry Applications, IEEE Transactions on,* vol. 48, pp. 191-198, 2012.
[6] K. Te-Tien, C. Chao-Shun, L. Chia-Hung, and H. Chin-Ying, "Design of phase identification system for phase measurement of distribution transformer," in *Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on*, 2012, pp. 1146-1149.
[7] H. Pezeshki and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network," in *Innovative Smart Grid Technologies (ISGT Europe), 2012 3rd IEEE PES International Conference and Exhibition on*, 2012, pp. 1-7.

[8] H. Pezeshki and P. Wolfs, "Correlation based method for phase identification in a three phase LV distribution network," in *Universities Power Engineering Conference (AUPEC), 2012 22nd Australasian*, 2012, pp. 1-7.
[9] M. H. F. Wen, R. Arghandeh, A. v. Meier, K. Poolla, and V. O. K. Li, "Phase Identification in Distribution Networks with Micro-Synchrophasors," presented at the IEEE Power and Energy Society General Meeting, Denver, 2015.
[10] B. K. Seal and M. F. McGranaghan, "Automatic identification of service phase for electric utility customers," in *Power and Energy Society General Meeting, 2011 IEEE*, 2011, pp. 1-3.
[11] T. A. Short, "Advanced Metering for Phase Identification, Transformer Identification, and Secondary Modeling," *Smart Grid, IEEE Transactions on,* vol. 4, pp. 651-658, 2013.
[12] V. Arya, D. Seetharam, S. Kalyanaraman, K. Dontas, C. Pavlovski, S. Hoy*, et al.*, "Phase identification in smart grids," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, 2011, pp. 25-30.
[13] M. Dilek, R. P. Broadwater, and R. Sequin, "Phase prediction in distribution systems," in *Power Engineering Society Winter Meeting, 2002. IEEE*, 2002, pp. 985-990 vol.2.
[14] P. Satya Jayadev, A. Rajeswaran, N. P. Bhatt, and R. Pasumarthy, "A novel approach for phase identification in smart grids using Graph Theory and Principal Component Analysis," in *2016 American Control Conference (ACC)*, 2016, pp. 5026-5031.
[15] V. Arya, V. T. Chakaravarthy, K. J. Dontas, S. T. Hoy, J. R. Kalagnanam, S. Kalyanaraman*, et al.*, "Systems and methods for phase identification," ed: Google Patents, 2014.
[16] "Smart Metering Implementation Programme Smart Metering Equipment Technical Specifications Version 1.58," Department of Energy and Climate Change, Ed., ed, 2014.
[17] Ofgem. (20/12/2015). *Transition to smart meters*. Available: https://www.ofgem.gov.uk/electricity/retail-market/metering/transition-smart-meters
[18] SEAI. (2012, 20/12/2015). *Press Release – Full Data from National Smart Meter Trial Published*. Available: http://www.seai.ie/News_Events/Press_Releases/2012/National_Smart_Meter_Trial_Data_Release.pdf

Minghao Xu received his B.Eng. degree in electrical and electronic engineering from the University of Bath, U.K, and electrical power engineering from North China Electric Power University, Baoding, China, in 2014. Currently, he is pursuing the Ph.D. degree at the University of Bath. His research interests include big data, machine learning and deep learning applications in power systems.

Ran Li received his B.Eng. degree in electrical power engineering from University of Bath, U.K, and North China Electric Power University, Beijing, China, in 2011. He received the Ph.D. degree from University of Bath, in 2014 and became a lecturer in Bath from 2015. His major interest is in the area of big data in power system, deep learning and power economics.

Furong Li (SM'09) was born in Shannxi province, China. She received the B.Eng. degree in electrical engineering from Hohai University, Nanjing, China, in 1990 and the Ph.D. degree from Liverpool John Moores University, Liverpool, U.K., in 1997. She is a professor in the Power and Energy Systems Group, University of Bath. Her major research interest is in the area of power system planning, analysis, and power system economics.